International Journal of Soft Computing and Artificial Intelligence, ISSN: 2321-404X, Volume-8, Issue-2, Nov-2020
http://iraj.in

USING BIG DATA TO COMPARE CLASSIFICATION MODELS FOR HOUSEHOLD CREDIT RATING IN KUWAIT

[1]NAJLA ALBARRAK,[2]HESSA ALSANOUSI, [3]IRENE MOULITSAS, [4]SALVATORE FILIPPONE

[1]Central Bank of Kuwait, P. O. Box: 526, Safat 13006, Kuwait
[2]Central Bank of Kuwait, P. O. Box: 526, Safat 13006, Kuwait
[3] Department of Computational Engineering Sciences, Cranfield University, Cranfield, Bedfordshire MK43 0AL, UK
[4]University of Rome Tor Vergata, Via Cracovia, 50, 00133 Roma RM, Italy
Email: [1]n.albarrak@cranfield.ac.uk, [2]h.alsanousi@cranfield.ac.uk, [3]i.moulitsas@cranfield.ac.uk,
[4]salvatore.filippone@cranfield.ac.uk,

Abstract - Credit rating risks have become the main indicator of bank performance. They are the reflection of the current status of the bank and an important milestone for future planning. An effective credit assessment can better anticipate expected losses and will minimize unexpected losses from accumulating. In an oil country such as Kuwait, advancements in technology as well as the big data available within banks about customers can lead to a built-in credit assessment model. This built-in model can work to help in-household credit scoring at the decision of a financial institution's management. Compared to the current 'black box' rating models, we did a comparison between different classification models for two types of banking: conventional and Islamic. The classification models are as follows: Logistic Regression, Fine Decision Tree, Linear Support Vector Machines, Kernel Naïve Bayes, and RUSBoosted. Sufficiently, the last could be used to classify banks household customers and determine their default cases.

Keywords - Classification Models, Conventional Banking, Credit Rating, Household Customers, Islamic Banking

I. INTRODUCTION

A financial crisis is not a new phenomenon or term. Rather, it is an ongoing bubble from the early stages of financial world. The early stages of financial development illuminated ways to avoid problems that could lead to a financial crisis. The main obstacle is how much this lesson costs and who oversees paying the bill. Banks in Kuwait have gone through several crises including the 2008 crisis and an impactful drop in oil prices. Kuwait stood strong without the need for any support from regulators for backup. This is due to the stringent regulations in place and the timely adoption of international regulations with additional safe buffers than internationally accepted benchmarks. Although Kuwait overcame the crisis, there is a space for improvement to ensure a more stable financial climate prior to a future financial. Crises cannot be fully avoided given the globalization in the business world; however, countries could introduce safeguards to be better prepared for them. After the 2008 international financial crisis, Basel committee emphasized on the minimum capital available to cover the riskiness of bank's assets and investment decisions.

The banking industry operates primarily through the usage of raised capital and borrowed funds to lend money and profit from the difference rates. Therefore, it should be noted that a bank's main activity is in lending money representing almost 60% of total banks assets in Kuwait [1]. From this amount, the share to household sector is growing for four consecutive years recording a growth of 7.8% in 2018 making it the top sector benefiting from credit growth in term of amount inFig. 1.

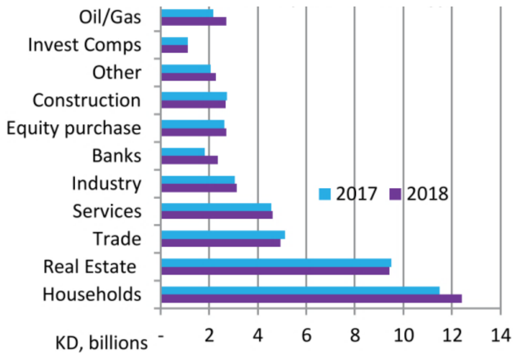

Figure 1: Gross Loans by Borrower's Type. Source: Central Bank of Kuwait, 2018

The increasing importance in household lending is due to the potential to distribute the risk of loan repayments to a wider range of customers. This is a contrast to big contracts with small number of corporations. On the other hand, excessive lending to one sector might increase the concentration risk which indicates more work is needed in analysing the riskiness to lend to household sectors. As a result, proposing a better credit scoring for the probability of default is a necessity in this case specially all customers (households, small/medium enterprises, and corporations) are uncategorized by riskeness in Kuwait and take the maximum risk weight due to that. In this paper, we propose a machine learning classification to categorize household credit default cases.
Recent events have shed importance on credit risks and drawn the attention of bankers and regulators with regards to managing the credit portfolio efficiently. There are several machine learning and deep learning options to examine credit probability

by default [2]. In a very recent International Monetary Fund working paper [3], the advantages in financial technology are discussed. Specifically, the literature investigates how machine learning solutions could reduce the cost of credit and to provide much clearer solutions than the 'black box' templates for nontechnical audiences. From corporate to household customers, the recent literature recognized the importance of programming in loan granting process.
Advancements in technology have created collaboration between the fields of computer science and finance. This collaboration leads to benefits as a result of the witnessed technological efficiency. In the context of using classification models for calculating the probability of customers loan default, there is a wide range of options to be implemented. Providing a credit rating model is fruitful for regulators and banks decision making evenly due to the advancement in the technology and the outperformance of the models developed and tested [4]. [5] compared 17 different methods for credit classes showing that it is suitable to conduct classification methods for credit rating. However, they did not use heavily imbalanced data in their case study. Imbalanced data in this context refers to data that is not well distributed between the two classes. In our data, almost 85% of the data is considered as good customers and the remaining 15% are prone to default. Given that when comparing machine learning methods to classify credit default customers, several studies showed that RUSBoost is the most significant method [6]. In [7], it is evidenced that that RUSBoost performance is significantly better for imbalanced data [8].

This paper aims to examine the gap of small size samples for studies with long-big data samples [2]. The literature recommends that ensemble learning (gradient boosting decision trees) is a solution for solving the disadvantages of decision trees specially if the data is large and has long history, which is the case of our research [3]. The research done on similar work was on a period of three years [4] while our aim is to expand it to five years minimum and up to 11 years. Another important aspect is that most studies have relayed on same set of data gathered from the customer disregarding the data available with banks. Our study relied on data available with the banks currently [9] and made use of it to estimate default customers. The paper will also provide a multi-period observation along historical events due to that the type of loan picked which has a term of 15 years. For the parameters, we chose customer characteristics already used along with new variables have not been used before: monthly number of transactions done in bank accounts and monthly average cash flow in bank accounts. Moreover, in order to evaluate our results properly, we used the standard measures in the field of credit scoring (Wang et al., 2012; Samreen, Zaidi and Sarwar, 2013). Specifically, the standard measures are average accuracy, type I error, and type II error.

II. DETAILS EXPERIMENTAL

2.1. Data

Our study is based on household customers in Kuwait with a sample of two banks: one conventional bank and one Islamic bank. The period from the data collection is, from 2008 to 2018, on a monthly basis. There are two types of household loans: consumer loans and installment loans.

- Consumer Loans: loans for the purpose of personal needs and durable goods with a limit of 15,000 KWD or 15 times the salary (whichever is less).
- Installment Loans: loans for the purpose of maintenance or purchase of private residents with a limit of 70,000 KWD.

Installment loans exposures were chosen given the higher amount granted which gives a higher impact to the economy. For Bank 1, the conventional bank, we took a sample of 100,000 customer base, out of which 37,488 have loans (installment and consumer). The number of customers with installment loans was 28,033 with 996 default cases. When we calculate this in terms of observations for machine learning classification, we have 347,977 transactions given that each customer could have more than one Installment loan. For Bank 2, the Islamic bank, we took a sample of 100,000 customer base, out of which 21,559 have loans (installment and consumer). The number of customers with installment loans are 15,108 with 1,394 default cases. When we calculate this in terms of observations for machine learning classification, we have 249,567 transactions given that each customer could have more than one installment loan.
To calculate the probability of household loans default, we gathered data of the loan's portfolio which included several items. The outstanding balance was gathered; this is the amount left from to be paid from the granted loan. The principal amount is a part of the monthly total payment to be paid against the principal amount of the loan. The remaining part of the monthly payment is the interest charge of the loan. Default cases are cases in which the customer fails to meet their monthly total obligations for three consecutive months while there is a remaining outstanding balance. This means that the customer is defaulting (bad customers). Hence, the ongoing payments of monthly obligations are considered non-defaulting (good customers).
The parameters, or independent variables, chosen for this study are:

- Credit card exposure
- Income
- Age
- Gender

- Education
- Nationality
- Relationship duration
- Monthly number of transactions done in bank accounts
- Monthly average cash flow in bank accounts
- Number of loans

A limitation with Bank 2 is that they did not provide us with nationality information.

2.2. Methodology

The aim of this study is to develop the most appropriate credit rating model to predict households default rate. In order to achieve this, we made a comparison between different classification model: Logistic Regression, Decision Tree, Linear Support Vector Machines, Bayesian Network, and RUSBoosted.

2.2.1. Logistic regression

In this paper, we will distinguish between two classes of creditors, good or bad [11]. For this binary response model, the response variable Y can take one of two set of values Y = 0 if the customer is good (non-defaulter), or Y = 1 if the customer is bad (defaulter). Xs are the columns vector of M explanatory variables, π = P (Y = 1|X) is the responses probability and N is the number of observations

$$\mathrm{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta^{T}x$$

where α is the intercept and βT is the coefficients.

2.2.2. Decision Tree

In classification decision trees, a single node is the starting point followed by binary differences (1,0). This results in the most information about the class [12]. Then, the process is repeated with the resulting new node until we reach a position to stop. Usually the tree is too large, so it is back-tested it through a cross-validation. The dependent variable Y is categorical, so, by using information theory in measuring how much we know about it from knowing the value of another separate variable A

$$I[Y;A] = \sum_{a} \Pr(A=a)\, I[Y;A=a]$$

I [Y ; A = a] is the value of the uncertainty about Y decreases from knowing that A = a given that we go from full population to sample where A = a. Therefore, I [Y ; A] is how much our doubt about Y reduces on average from knowing the value of A.

2.2.3. Support Vector Machines

From assuming a training set of N {(Xi, Yi)}Ni=1 with input data Xi ε Rn and consistent binary class labels Yi ε {-1+1}, the SVM classifier in Vapnik's theory satisfies the following:

$$yi[w^{T}\varphi(xi) + b] \geq 1, \quad i = 1, \dots, N$$

The non-liner function of φ (.) plots the input space to a high dimensional feature space [5]. In this space, the mentioned variations construct a hyperplane WT φ (X) + b = 0 discriminating between two classes. In the original weighted space, the following equation is used for the classifier:

$$y(x) = \mathrm{sign}[w^{T}\varphi(x) + b]$$

However, it is never evaluated in this form where the curved optimization problem could be defined as

$$\min_{w,b,\xi} j\,(w,b,\xi) = \frac{1}{2}w^{T}w + C\sum_{i=1}^{N}\xi i$$

subject to

$$\begin{cases} yi[w^{T}\varphi(x_i) + b] \geq 1 - \xi_i, & i = 1, \dots, N \\ \xi_i \geq 0, & i = 1, \dots, N \end{cases}$$

The variables used in ξi are loose variables which are needed to allow the misclassifications to occur in the set of inequalities due to overlying distribution. The first section of the objective function is set to maximize the margin between two classes in the feature space. The second part is set to minimize the misclassification error.

2.2.4. Bayesian Network

Bayesian Network is a simple and high performance classifier (Baesens et al., 2003). This classification model works through learning the class condition probability p (Xi|Y) from each input variable Xi I = 1 … n given the class label Y. Then, a new observation is classified by Bayes' rule to calculate the following probability of each class of Y given the vector of observed feature values:

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

To make things easier, an assumption behind the naïve Bayes classifier is that the features are in theory independent given the class label, therefore

$$p(x|y) = \prod_{i=1}^{n} p(x_i|y)$$

The probabilities p (Xi|Y) are then estimated through using the frequency counts for the discrete features and a normal based method for the continues features.

2.2.5. Gradient boosting

Gradient boosting technique, which is an ensemble algorithm founded by [13], is used to calculate the probability of loan default. It relies on incremental minimization of the error term which improves the precision of the prediction function [11]. In trees, and after setting the learner base, every tree calculated is fit to the 'pseudo residuals', which is the deviation from the median and not from the expectation, from the earlier predictions in order to lower the error in general. Therefore, the following model is used:

$$F(x) = G_0 + \beta_1 T_1(x) + \beta_2 T_2(x) + \cdots + \beta_n T_n(x)$$

G0 is the initial value for the set. T1 … Tn are the trees and β1 … βn are the coefficients for particular tree nodes calculated by the algorithm. To conduct the gradient boosting classifier, a maximum branch size needs to be set. For this study, 30 learning cycles and a 0.1 learn rate were set.

The selection of the appropriate model will be through an evaluation for the Area Under Curve (AUC) for the Receiver Operating Characteristic (ROC) curve in Fig 2[14]. To better interpret these results, we could define the ROC as a curve connecting more predictive data points starting from less than the lowest value observed, (0,0), and ends at greater than the highest value observed (1,1).

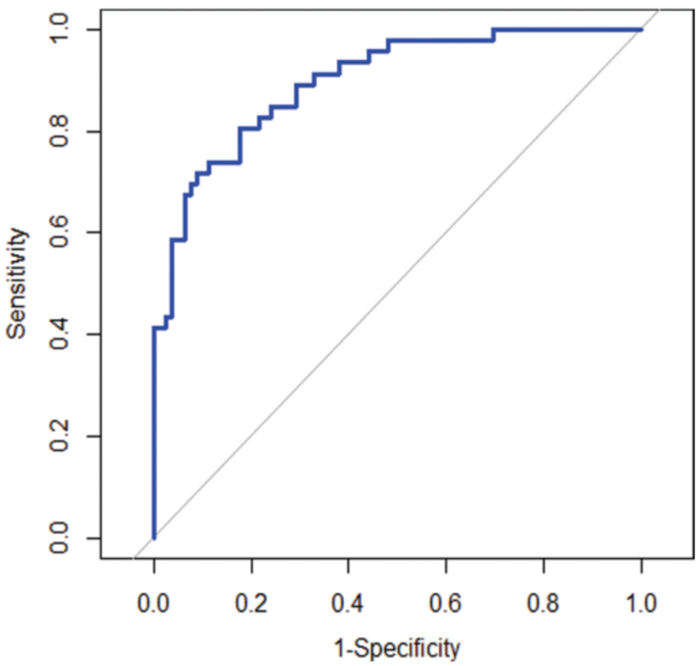

Figure 2: Hypothetical ROC curve

After drawing the ROC curve, the AUC is the entire area underneath the ROC curve. The diagonal line between points (0,0) and (1,1) indicates that the values on this line are not indicative of a better estimate than a random guess (AUC = 0.50). The further the point in the ROC space above the diagonal line, the better the predictive value of the test.

Next, an overlook for the true positive rates and false negative rates, from the confusion matrix, is conducted to enhance our analysis. Through the information developed in the confusion matrix in Table 1, we will construct our analysis on calculating the accuracy rate, type I error and type II error.

		Actual Condition	
		Positive (non-risk)	Negative (risk)
Test	Positive (non-risk)	True Positive (TP)	False Positive (FP)
Result	Negative (risk)	False Negative (FN)	True Negative (TN)

Table 1: Confusion matrix for credit scoring

The three measures are defined as

$$\text{Averageaccuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}$$

$$\text{TypeIError} = \frac{\text{FN}}{\text{TP} + \text{FN}}$$

$$\text{TypeIIError} = \frac{\text{FP}}{\text{TN} + \text{FP}}$$

III. RESULTS AND DISCUSSION

After running the classification models to predict the probability of default, a comparison of the AUC is done to facilitate which model to choose inTable 2.

Bank 1	AUC	Bank 2	AUC
Logistic Regression	0.71	Logistic Regression	0.72
Decision Tree	0.7	Decision Tree	0.69
Linear Support Vector Machines	0.45	Linear Support Vector Machines	0.66
Bayesian Network	0.67	Bayesian Network	0.70
RUSBoosted	0.86	RUSBoosted	0.80

Table 2: AUC results

Table 3 and 4presents a summary of the confusion matrix output and the performance indicators.

Bank 1	True Positive	True Negative	False Positive	False Negative	Average Accuracy	Type I Error	Type II Error
Logistic Regression	100%	0%	0%	100%	50%	50%	0%
Decision Tree	99%	0%	1%	100%	50%	50%	1%
Linear Support Vector Machines	100%	0%	0%	100%	50%	50%	0%
Bayesian Network	99%	0%	1%	100%	50%	50%	1%
RUSBoosted	84%	74%	16%	26%	79%	24%	16%

Table 3: Confusion matrix results and model; performance results for Bank 1

Bank 2	True Positive	True Negative	False Positive	False Negative	Average Accuracy	Type I Error	Type II Error
Logistic Regression	99%	0%	1%	100%	50%	50%	1%
Decision Tree	99%	2%	1%	98%	51%	50%	1%
Linear Support Vector Machines	100%	0%	0%	100%	50%	50%	0%
Bayesian Network	99%	7%	93%	1%	53%	1%	48%
RUSBoosted	67%	78%	22%	33%	73%	33%	25%

Table 4: Confusion matrix results and model; performance results for Bank 2

From the illustration provided earlier, the RUSBoosted can be the most efficient model for calculating the probability of default. This is attributed to the large sample set of Bank 1 that runs for a long period of 11 years. This huge data sets shows the lack of existing classification methods and enhance the importance of ensemble models. From an AUC of 0.86 to average accuracy of 79% with lowest value in type I error (24%), the RUSBoosted is the ideal solution for big data classifications especially in imbalanced data. The increase rate in type II error of 16% could be justified by that RUSBoosted is the only model working for the data set on hand indicating false cases while the rest models are overfitted with 0% or 1% values in false positive. An important clarification is that this model is used to estimate the credit default for household's customers in order to classify the customers in to stages for determining the unexpected loss of good customers. Given that we defined the default rate for three consecutive non-payments, as per the provisioning scheme, the high false-negative rates are covered risk wise through the provision charges.

In Bank 2, an AUC of 0.80 and average accuracy of 73% with lowest value in type II error, 25%, the RUSBoosted is the ideal solution for big data classifications especially in imbalanced data. The increase rate in type I error of 33% could be justified by, as stated, that RUSBoosted is the only model working for the data set on hand indicating false cases while the rest models are overfitted with zero or 1% values in false positive.

To evaluate our model, a training subsample for bank 1 of 70% (243,584 observations) from total observations of 347,977 has been tested and the following AUC has been calculated to show the accuracy of the model in Fig. 3. Nevertheless, a training subsample for Bank 2 of 70% (174,697 observations) from total observations of 249,567 has been tested and the following AUC has been calculated to show the accuracy of the model in Fig. 4.

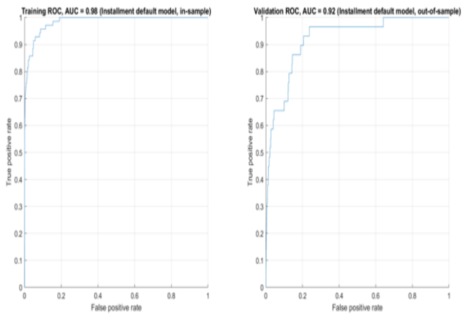

Figure 3: AUC results for training and validation Bank 1

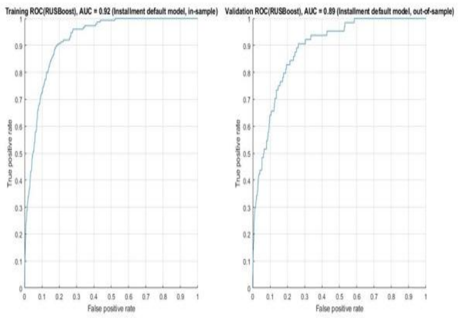

Figure 4: AUC results for training and validation Bank 2

For our parameters used in Bank 1, Fig. 5 shows the relative importance of each parameter indicating that 'Income' is the most influenced with 36.1% reliance for predictions. This is followed by 'Nationality', 'Relationship duration' and 'Credit card exposure' with weights of 15.1%, 11.3%, and 9.5% respectively. The results show relative importance of new variables selected for the first time 'Monthly average cash flow' and 'Monthly number of transactions' with weights of 7% and 5% respectively. This is an indication that internal information held within banks about the customers is important for both credit evaluation and prediction. For our parameters used in Bank 2,Fig. 6 shows the relative importance of each parameter indicating that 'Number of loan counts' is the most influencer with 30.8% reliance for predictions. This is followed by 'Income', 'Monthly average cash flow' and 'Relationship duration' with weights of 19.4%, 10.9%, and 10.4% respectively. This fact supports the significance of our theory, especially for Islamic banks.

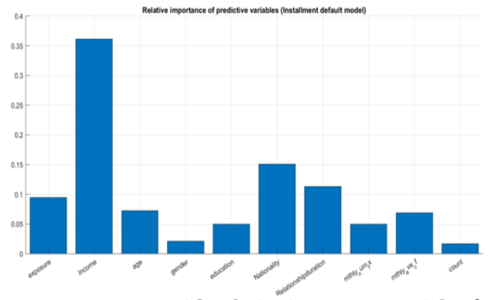

Figure 5: Parameters with relative importance weights for Bank 1

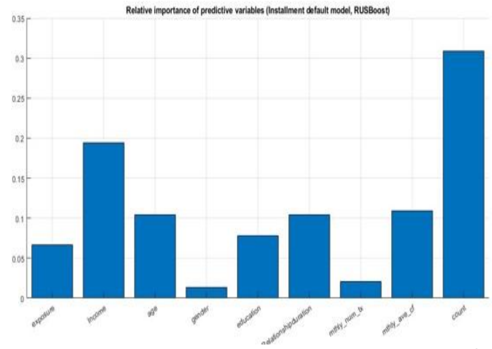

Figure 6: Parameters with relative importance weights for Bank 2

The results indicate that the new introduced parameters can have a high impact in predicting probability of installment household loans, specifically monthly average cash flow and relationship duration between the bank and the customers in Kuwait. Knowing customers habits regarding withdrawals and deposits on monthly basis is a clue to their behavior after loans have been granted. Additionally, loyalty could be inferred by the long relationship duration where customers tempt to repay their loans smoothly if they have long relationship duration with the bank and vice versa. This will shift the view in creating the database of inputs from customers' given information to including inhouse data about the client within the bank. This information is suitable for both type of banking, conventional and Islamic. This research asserts that it is not only important for conventional banks in Kuwait to have an internal credit rating model, but it also it is suitable for Islamic banking as well.

IV. CONCLUSION

In general, classification methods are increasingly implemented in fields other than computer science. The literature review is full of studies evidencing the efficiency of such models in knowing the expected resulting different classes. Nevertheless, classification methods have been used to categorize credit default classes, good or bad, in order to prepare regulators and bankers to better anticipate risks. This paper compared the different methods of classifications (Logistic Regression, Decision Tree, Linear Support Vector Machines, Bayesian Network and RUSBoosted) in order to examine the credit default cases. Our study relied on big data from Kuwaiti banks for 11 years to tackle the gap of not lengthy data. The parameters also have included new items -- more than what was used previously. Moreover, those data came from banks indicating the importance of data existing within banks data bases. From the AUC, average accuracy, type I error, and type II error RUSBoosted were chosen as the outperforming method. A supporting result for training sets have indicated the efficiency of the model selected. From the relative importance weight, we can identify the important parameters used for other banks in order to obtain information for their studies.

ACKNOWLEDGMENTS

This work is sponsored by the Central Bank of Kuwait.

REFERENCE

[1] Central Bank of Kuwait: Financial Stability Report 2018. Cent. Bank Kuwait. 1–69 (2018). https://doi.org/10.1007/s13398-014-0173-7.2.
[2] Wang, G., Ma, J., Huang, L., Xu, K.: Two credit scoring models based on dual strategy ensemble trees. Knowledge-Based Syst. 26, 61–68 (2012). https://doi.org/10.1016/j.knosys.2011.06.020.
[3] Bazarbash, M.: FinTech in Financial Inclusion: Machine Learning Applications in Assessing Credit Risk. IMF Work. Pap. 19, 1 (2019). https://doi.org/10.5089/9781498314428.001.
[4] Petropoulos, A., Siakoulis, V., Stavroulakis, E., Klamargias, A.: A robust machine learning approach for credit risk analysis of large loan level datasets using deep learning and extreme gradient boosting. use big data Anal. Artif. Intell. Cent. Bank. 50, 30–31 (2018).
[5] Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., Vanthienen, J.: Benchmarking state-of-the-art classification algorithms for credit scoring. J. Oper. Res. Soc. 54, 627–635 (2003). https://doi.org/10.1057/palgrave.jors.2601545.
[6] Munkhdalai, L., Munkhdalai, T., Namsrai, O.E., Lee, J.Y., Ryu, K.H.: An empirical comparison of machine-learning methods on bank client credit assessments. Sustain. 11, 1–23 (2019). https://doi.org/10.3390/su11030699.
[7] Seiffert, C., Khoshgoftaar, T.M., Van Hulse, J., Napolitano, A.: RUSBoost: A hybrid approach to alleviating class imbalance. IEEE Trans. Syst. Man, Cybern. Part ASystems Humans. 40, 185–197 (2010). https://doi.org/10.1109/TSMCA.2009.2029559.
[8] Seiffert, C., Khoshgoftaar, T.M., Van Hulse, J., Napolitano, A.: RUSBoost: Improving classification performance when training data is skewed. Proc. - Int. Conf. Pattern Recognit. (2008). https://doi.org/10.1109/icpr.2008.4761297.
[9] Nyathi, K., Ndlovu, S., Moyo, S., Nyathi, T.: Optimisation of the Linear Probability Model for Credit Risk Management. Int. J. Comput. Inf. Technol. 03, 1340–1345 (2014).
[10] Samreen, A., Zaidi, F.B., Sarwar, A.: Design and development of credit scoring model for the commercial banks of Pakistan: forecasting creditworthiness of individual borrowers. Int. J. Bus. Soc. Sci. 2, 1–26 (2013).
[11] Brown, I., Mues, C.: An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Syst. Appl. 39, 3446–3453 (2012). https://doi.org/10.1016/j.eswa.2011.09.033.
[12] Speybroeck, N.: Classification and regression trees. Int. J. Public Health. 57, 243–246 (2012). https://doi.org/10.1007/s00038-011-0315-z.
[13] Friedman, J.H.: Stochastic gradient boosting. Comput. Stat. Data Anal. 38, 367–378 (2002). https://doi.org/10.1016/S0167-9473(01)00065-2.
[14] Yang, S., Berdine, G.: The receiver operating characteristic (ROC) curve. Southwest Respir. Crit. Care Chronicles. 5, 34 (2017). https://doi.org/10.12746/swrccc.v5i19.391.

★★★

2021-03-13

Using Big Data to compare classification models for household credit rating in Kuwait

Albarrak, Najla

Inderscience

Albarrak N, Alsanousi H, Moulitsas I, Filippone S. (2020) Using Big Data to compare classification models for household credit rating in Kuwait. International Journal of Soft Computing and Artificial Intelligence, Volume 8, Issue 2, November 2020, pp. 1-7

http://iraj.doionline.org/dx/IJSCAI-IRAJ-DOIONLNE-17674